BROADCOM

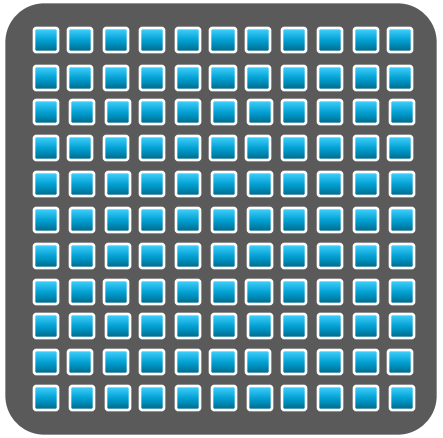
AI Networking - Requirements & Solutions

Kamran Naqvi

Chief Network Architect – EMEA, CSG

October 2025

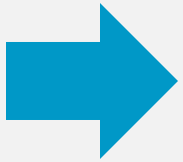
AI … A Very Very Large Distributed Computing System

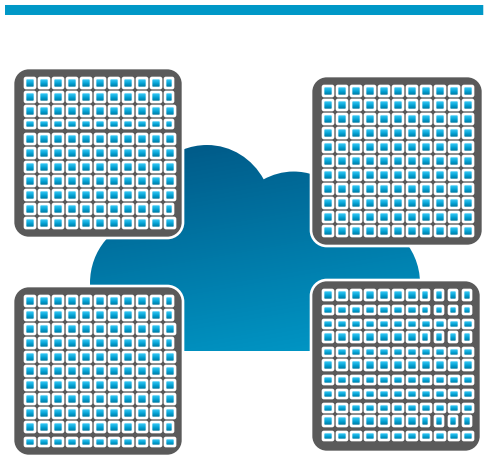


GPU

Optimized for
Parallel Tasks

Thousands of Cores





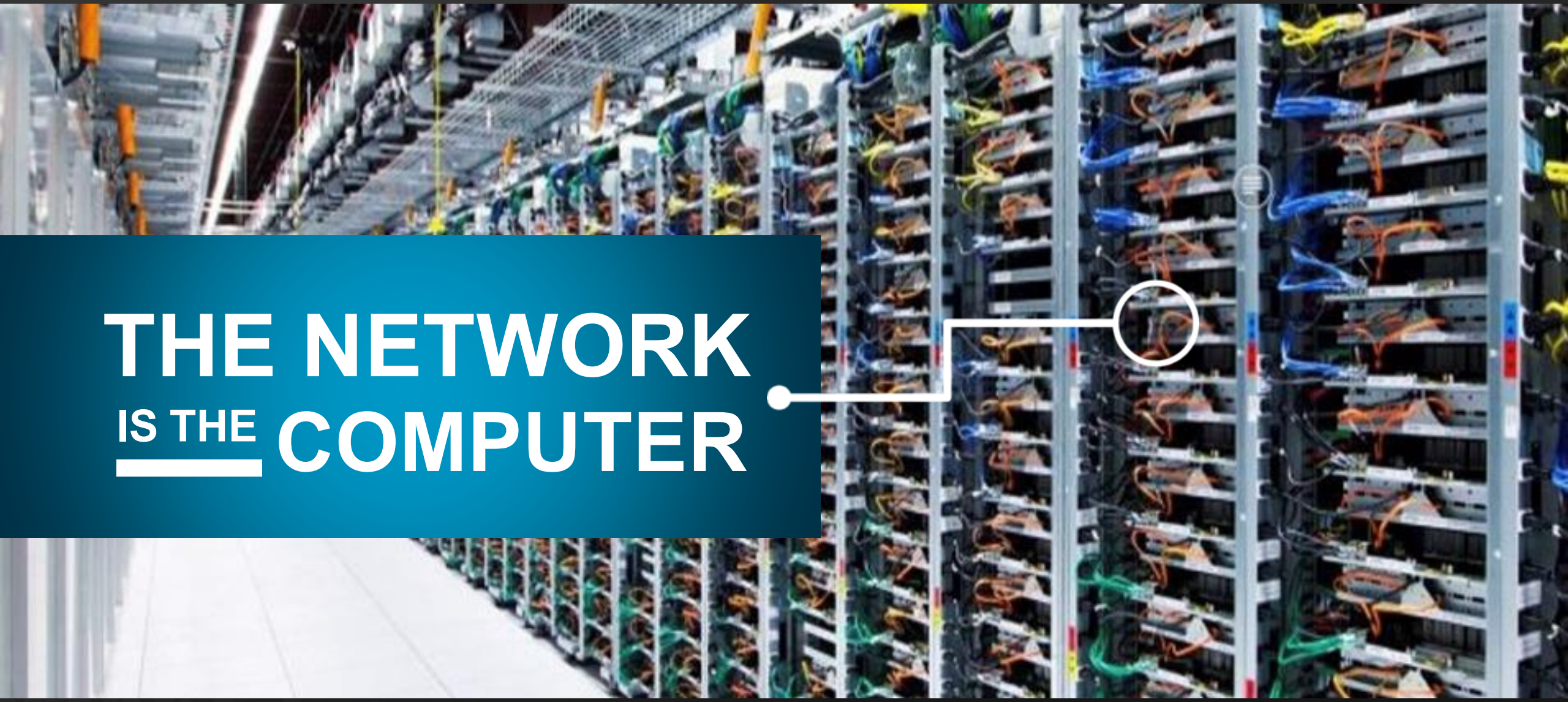
GPU Clusters

GPU Network for AI
Workloads

100K+ GPUs

OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

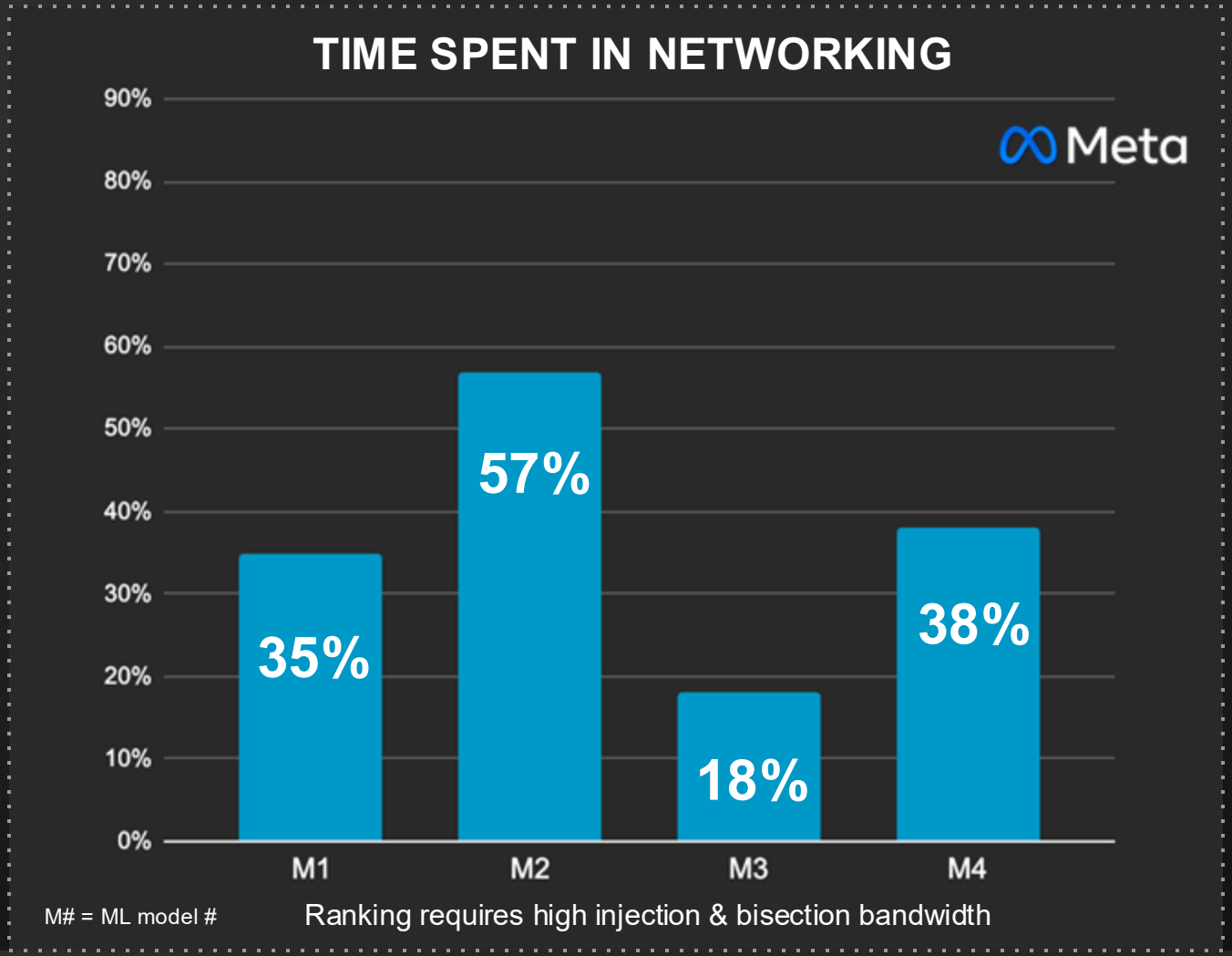


OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

“Network I/O is Key for Recommendation Workloads.”

OCP keynote by Alexis Bjorlin
at 2022 OCP Global Summit

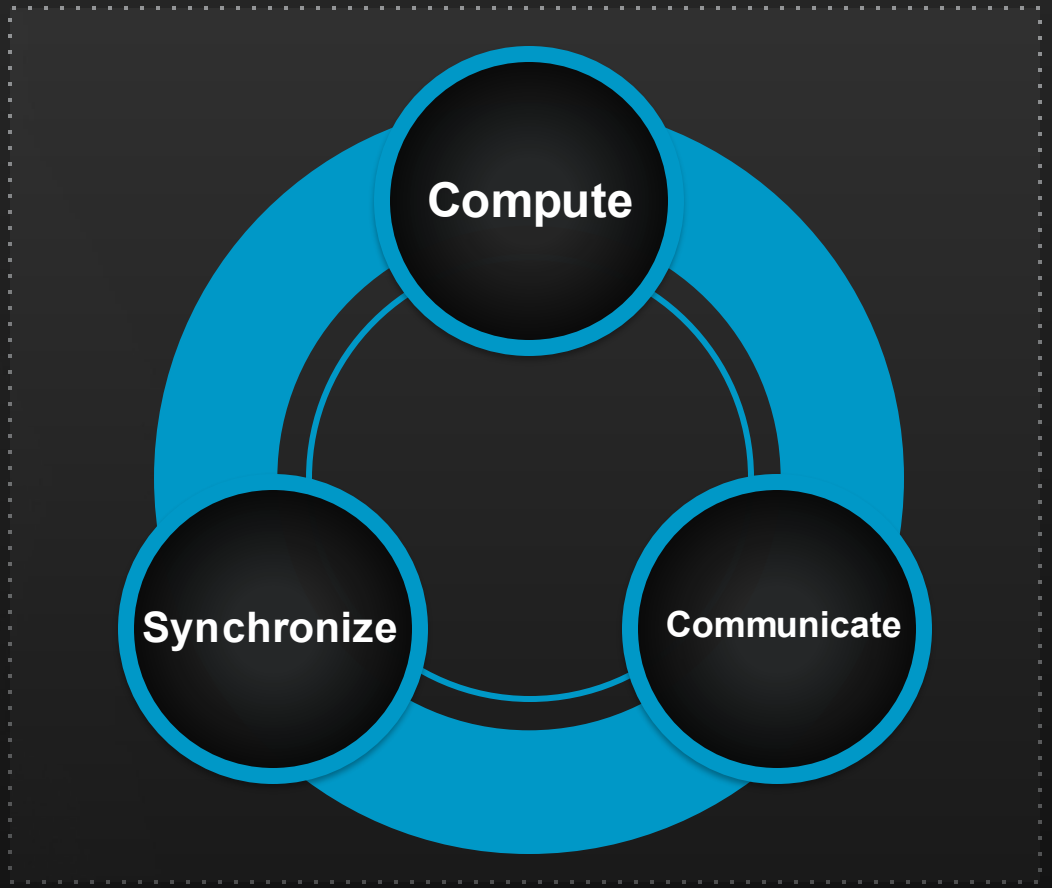


OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

What Makes AI Networks Unique?

- High bandwidth
- Elephant flows
- Synchronized and bursty traffic
- RDMA dominant traffic
- Training jobs run for long periods of time (hours, days)
- Tail latency impacts job completion time significantly

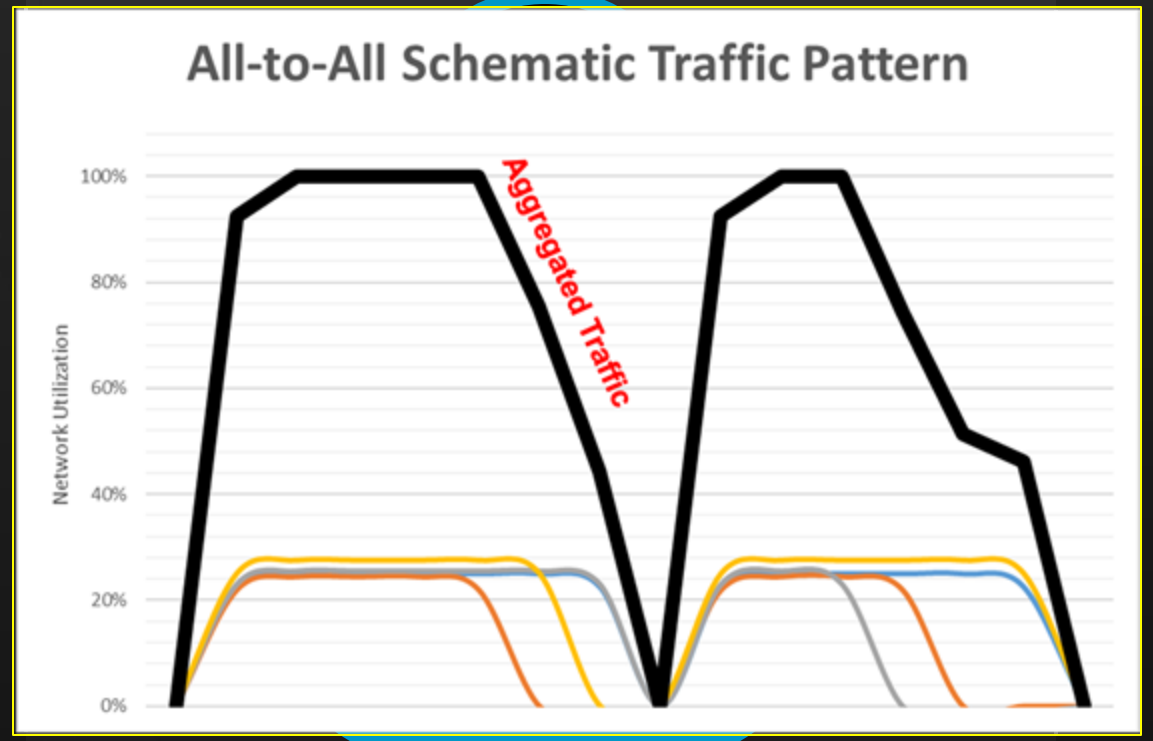


OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

What Makes AI Networks Unique?

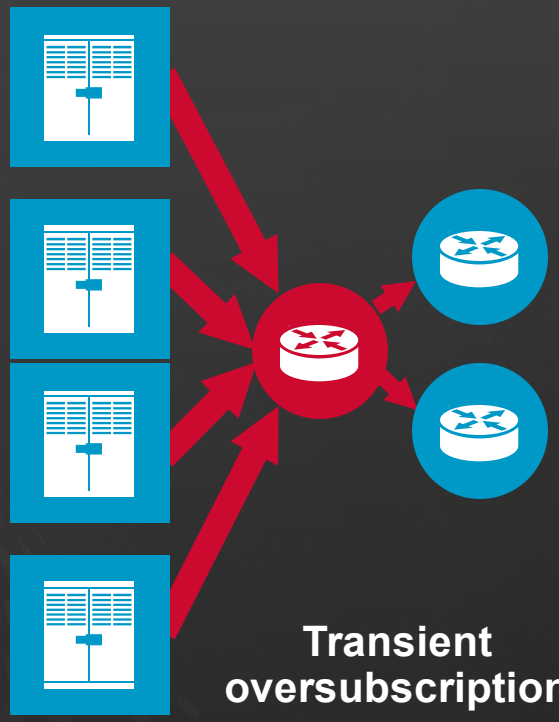
- High bandwidth
- Elephant flows
- Synchronized and bursty traffic
- RDMA dominant traffic
- Training jobs run for long periods of time (hours, days)
- Tail latency impacts job completion time significantly



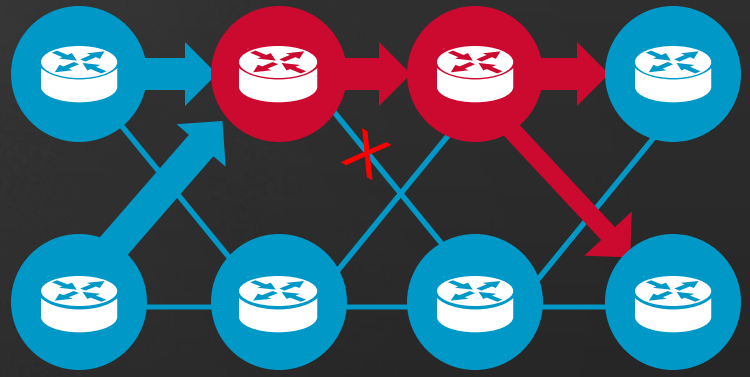
OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

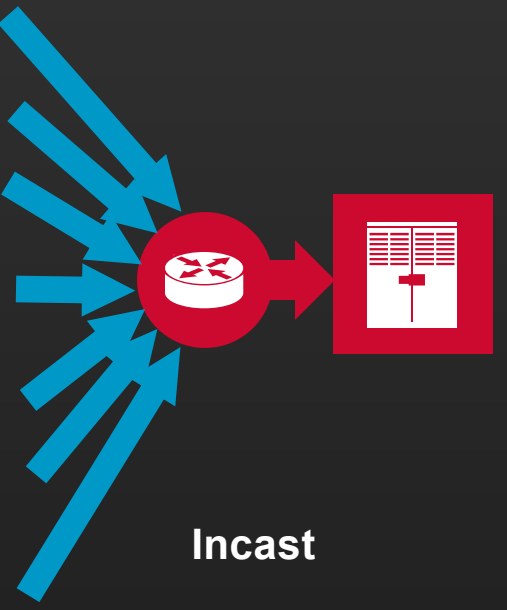
“Time Spent in Networking” is Impacted By…



Transient oversubscription

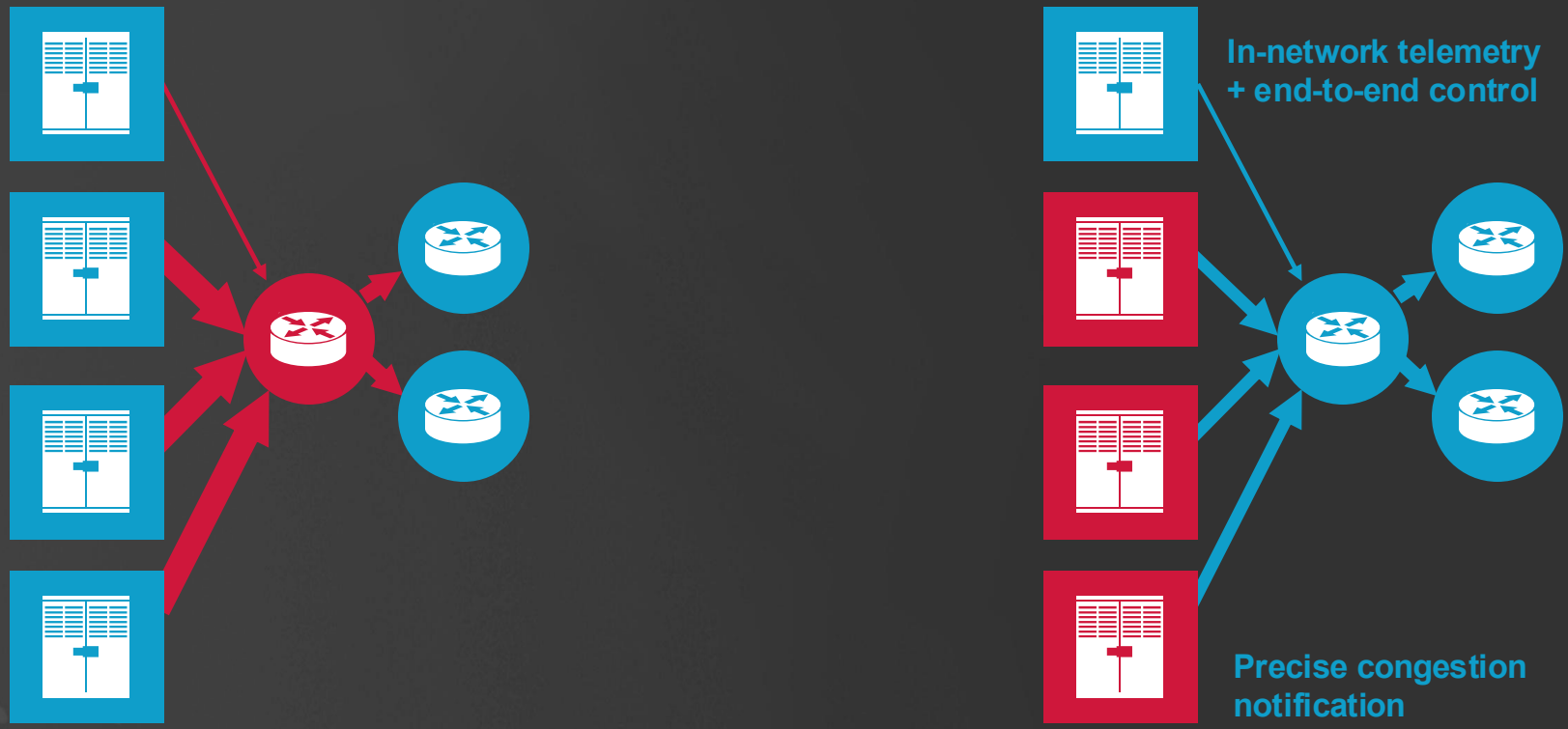


Flow collisions & Link failures



Incast

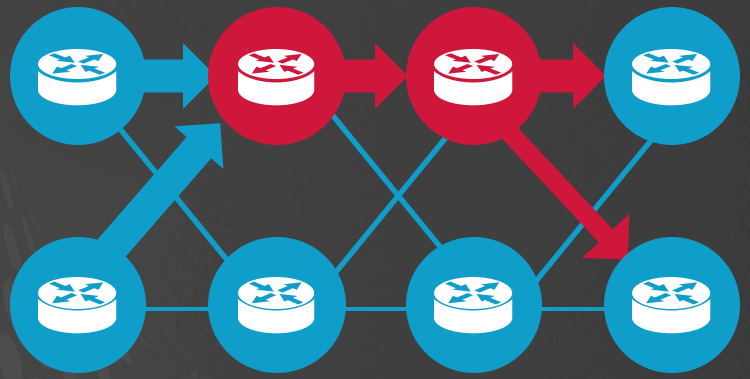
“Time Spent in Networking” is Improved By…



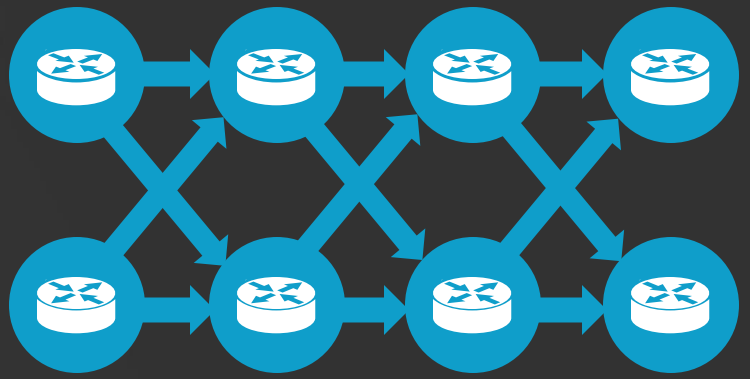
OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

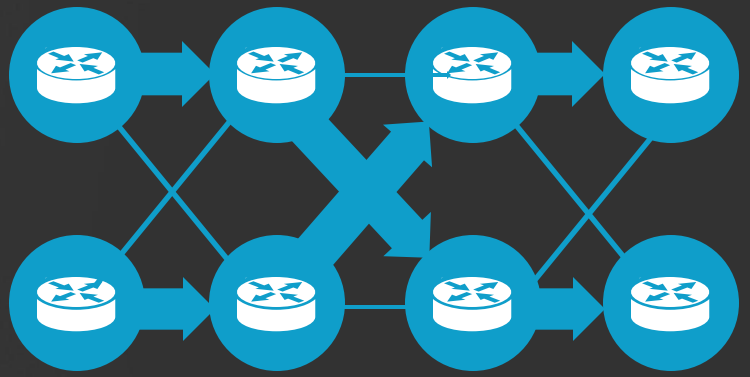
“Time Spent in Networking” is Improved By…



Packet spraying with receiver ordering

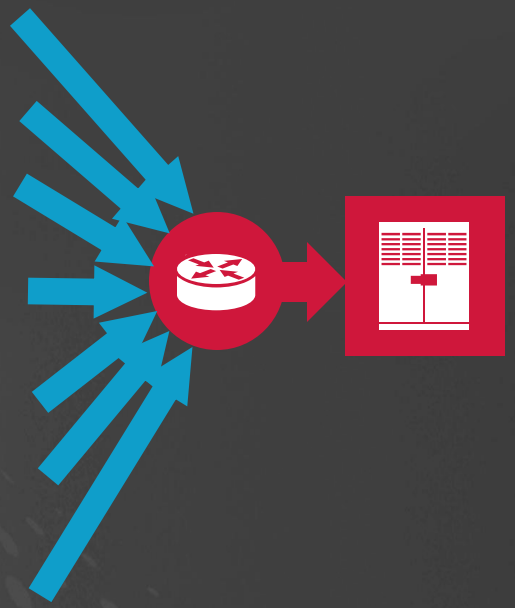


Cognitive routing (load-aware ECMP)



In case of link failure recovery should happen in hardware … Zero Impact Failover (ZIF)

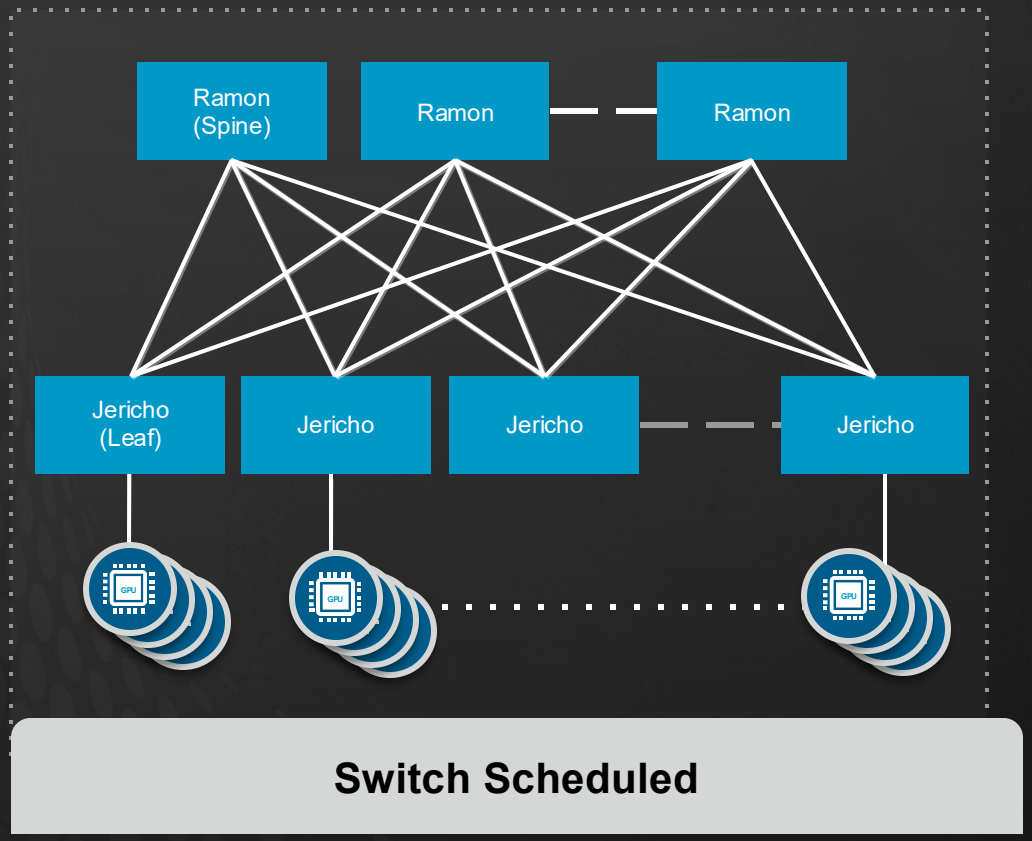
“Time Spent in Networking” is Improved By…

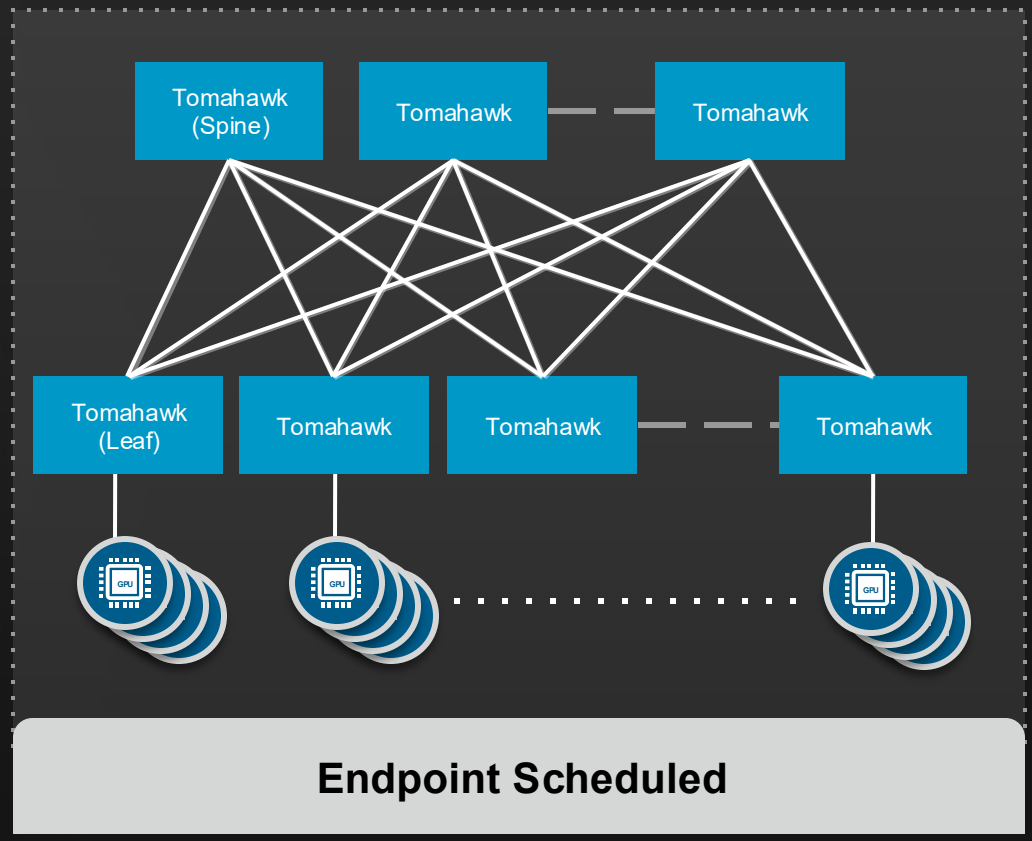


Receiver-based credit control can pace senders accurately.

Credit control mechanism can exist on the switch or the endpoint

Broadcom's AI Fabric Solutions





OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

Tomahawk 6: Built for AI Scale

World's First 102.4 Tbps Switch Chip

- Double the bandwidth of any other Ethernet switch
- Built to power clusters with 1M+ XPUs

Unmatched Performance & Power Efficiency

- Cognitive Routing 2.0, Deep Insight
  Advanced 3nm technology

Unrivaled Versatility

- Scale-up & scale-out, training & inference
- Works with any endpoint, including XPU scale-up interfaces

Industry-Leading SerDes and CPO

- Options for 512x200G PAM4, 1024x100G PAM4, CPO
- Ground-breaking SerDes and optics density

NOW SHIPPING

Tomahawk6-200G

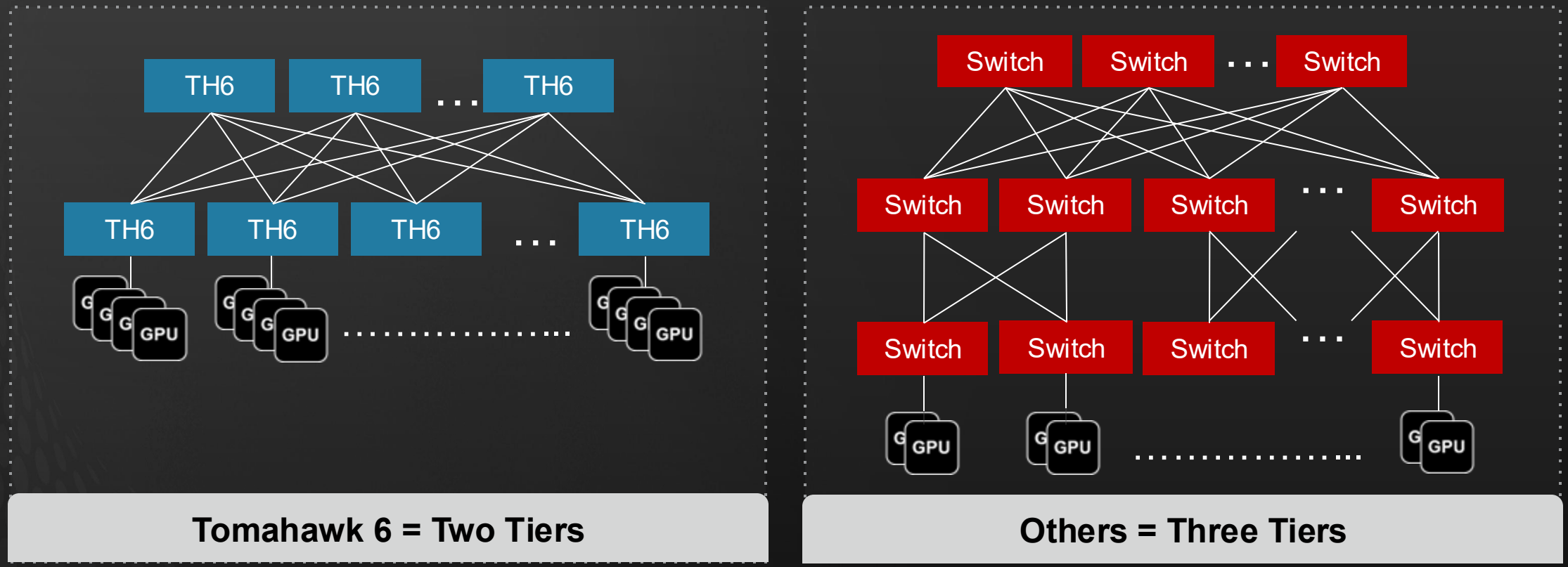
Tomahawk6-100G

Support for 1.6TbE Ports
Ultra Ethernet Compliant

OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

Tomahawk 6: Large Two-Tier Scale-Out



N = # rails

128K-GPU Cluster at 200GbE x N per Endpoint

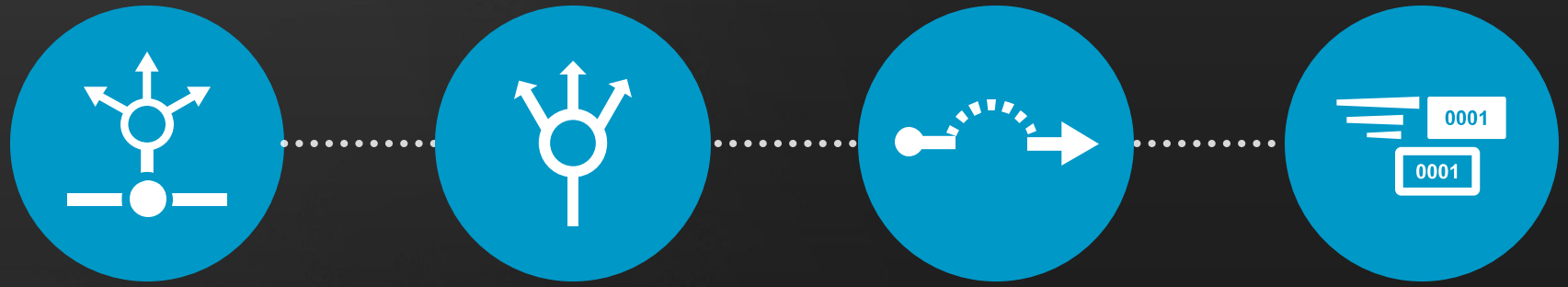
OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

Benefits of Two-Tier vs. Three-Tier Networks

Two-Tier	Three-Tier
Fewer optics	67% more optics
Lower latency	67% higher latency (5 vs. 3 hops)
Higher reliability	>3x number of switches & 67% more optics
Higher performance	Difficult load balancing and congestion control, more link flaps
Lower power	2x power for the network

Broadcom Cognitive Routing 2.0



Global Load Balancing

Egress link selection based on global path congestion

Reactive Path Rebalancing

Update egress links for active flows when congestion is detected

Fast Link Failover

Automatically steers traffic around failed links in <500ns

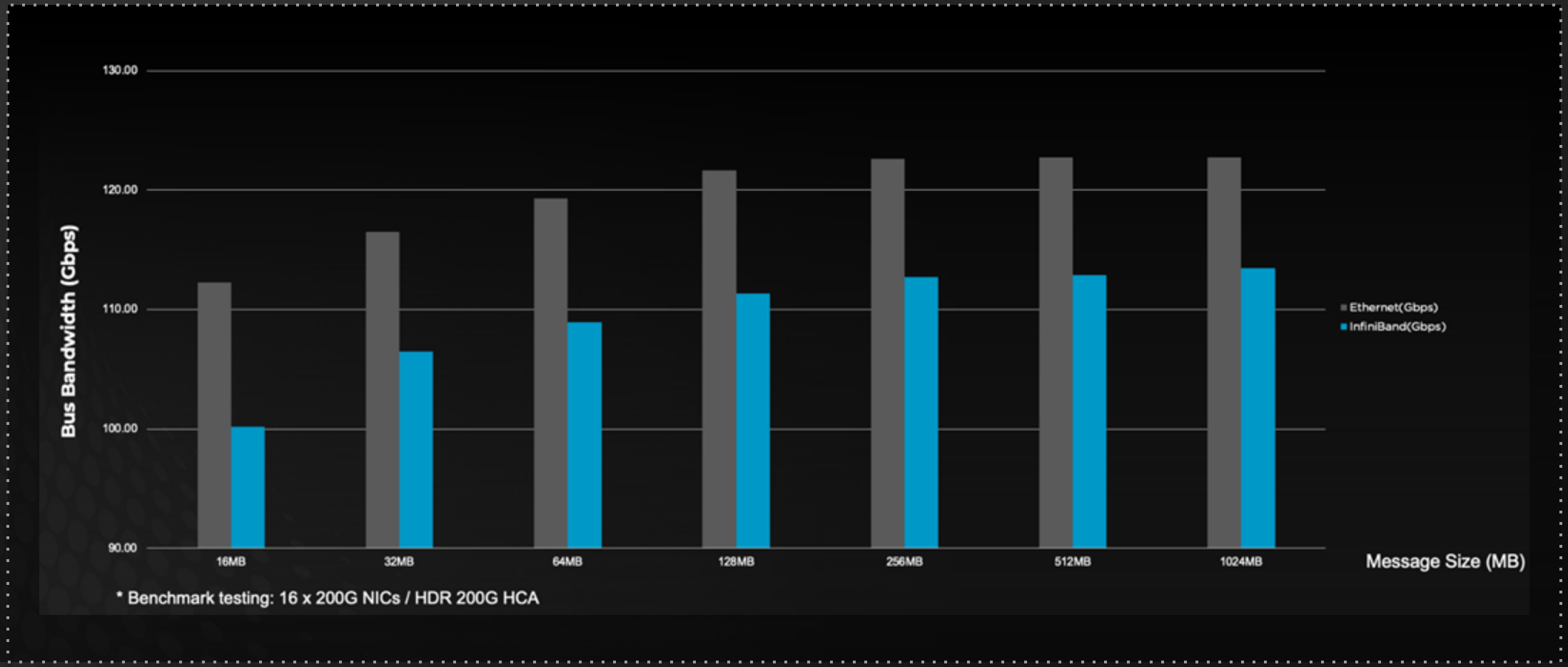
Drop Congestion Notification

For congested queues, send trimmed packet at high priority to destination

New for Cognitive Routing 2.0:

- Increased scale
- Faster and enhanced congestion signaling
- Enhanced path selection algorithms

Ethernet vs. InfiniBand: 10+% Improvement in Job Completion Time



BROADCOM

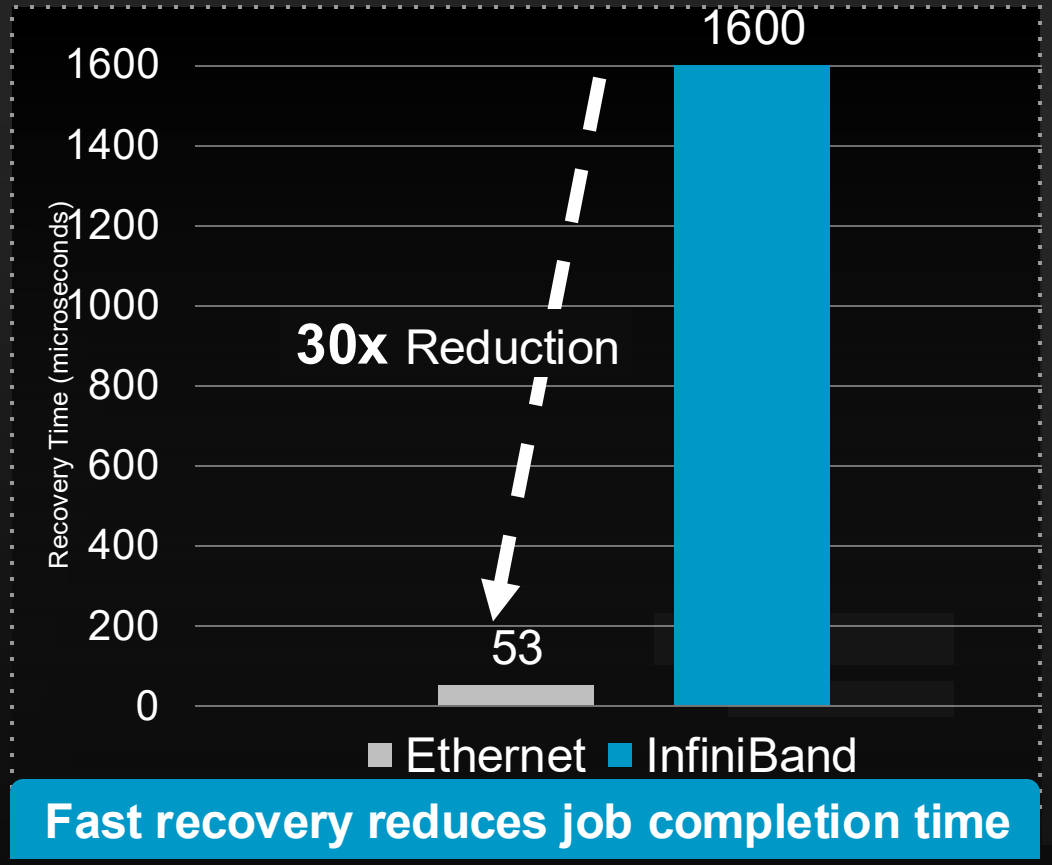
Ethernet Provides 30x Faster Failover than InfiniBand

2%

Optics Annual Failover Rate*

15

Failures per Month**



Fast recovery reduces job completion time

* Typical industry failure rate. ** Assuming 4K node cluster using 9.2K optic modules

OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

All Hyperscalers: Ethernet AI fabric

Alibaba Group 阿里巴巴集团	100,000+
amazon	60,000+
ORACLE	30,000+
Meta	30,000+
ByteDance	10,000+

Baidu 百度

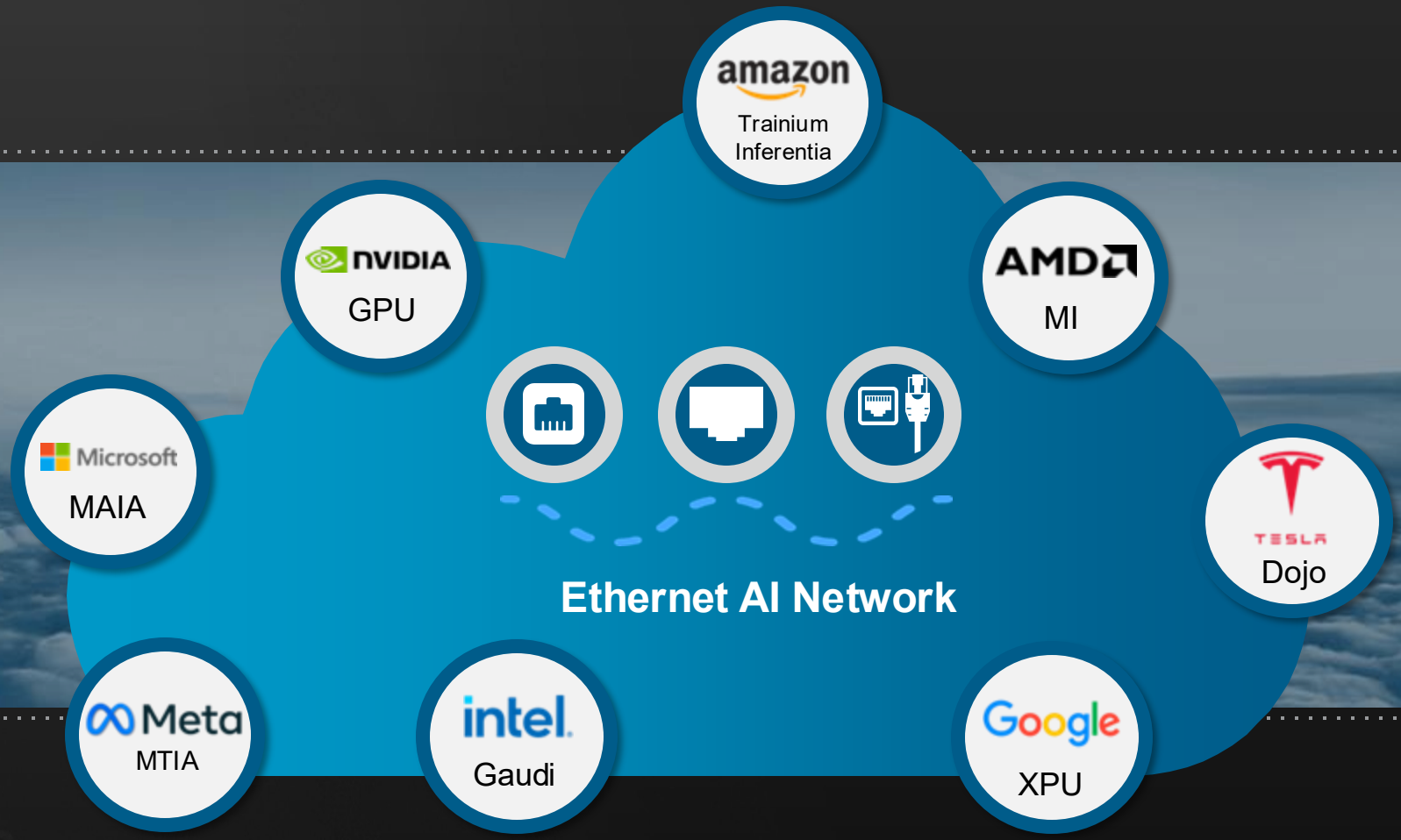
Tencent 腾讯

Microsoft

OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

Ethernet Network for AI



OPEN // SCALABLE // POWER EFFICIENT

BROADCOM

BROADCOM®

Thank You